DOCUMENT RESUME

ED 067 649 CS 000 185

AUTHOR	Cooper, Charles R.
TITLE	Measuring Growth in Appreciation of Literature. Reading Information Series: Where Do We Go?
INSTITUTION	Indiana Univ., Bloomington. ERIC Clearinghouse on Reading.; International Reading Association, Newark, Del.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	72
NOTE	30p.
AVAILABLE FROM	International Reading Association, Six Tyre Avenue, Newark, Del. 19711 ($1.00 member, $1.50 nonmember)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Affective Behavior; Content Analysis; *Literature Appreciation; *Poetry; *Prose; *Reading Research; Validity

ABSTRACT

This monograph is written primarily for the researcher. It reviews a number of attempts to measure appreciation of literature. The measurements are grouped in two categories: (1) discrimination among poems or prose extracts, and (2) content analysis. Following the review is an evaluation of the limitations and possibilities of these measures. The monograph concludes with specific recommendations for further research into the problem of measuring growth in appreciation of literature. (Author)

ED 067649

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Measuring Growth in Appreciation of Literature

Charles R. Cooper
State University of New York at Buffalo

ec+ira
Reading Information Series: WHERE DO WE GO?
1972

International Reading Association
Six Tyre Avenue
Newark, Delaware 19711

This series from ERIC/CRIER+IRA is designed to review the past, assess the present, and predict the future. This paper reflects the continued careful and thoughtful development of the series by Dr. Richard A. Earle.

James L. Laffey
Director of ERIC/CRIER

The International Reading Association attempts, through its publications, to provide a forum for a wide spectrum of opinion on reading. This policy permits divergent viewpoints without assuming the endorsement of the Association.

Measuring Growth in Appreciation of Literature was prepared pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.

Contents

4 Foreword

5 Introduction

8 Review of Attempts to Measure Appreciation

- Discrimination Among Prose Extracts and Poems
- Content Analysis

15 Synthesis

- Discrimination Among Poems and Prose Extracts
- Content Analysis

18 Recommendations

- Specific Recommendations

22 Bibliography

Foreword

ERIC/CRIER and IRA are concerned with several types of information analysis and their dissemination to audiences with specific professional needs. Among these is the producer of research—the research specialist, the college professor, the doctoral student. It is primarily to this audience that the present series is directed, although others may find it useful as well. Therefore, the focus will rest clearly on the *extension* of research and development activities: "Where do we go?" Our intent is not to provide a series of exhaustive reviews of literature. Nor do we intend to publish definitive statements which will meet with unanimous approval. Rather, we solicit and present the thoughtful recommendations of those researchers whose experience and expertise have led them to firm and well-considered positions on problems in reading research.

The purpose of this series of publications is to strengthen the research which is produced in reading education. We believe that the series will contribute helpful perspectives in the research literature and stimulating suggestions to those who perform research in reading and related fields.

Richard A. Earle
Series Editor

Introduction

Obtaining valid and reliable measures of all forms of student learning and behavior has always been a challenging task for the schools. In literary study, where the most important goals are those of response, value, and discrimination, the problem of measurement is particularly acute. It has been concisely stated: "What we want to measure is complex but subjective; the methods we have to work with are objective but simple. The problem, then, is to make our goals more objective and our measures more complex" (Forehand, 1966).

It is important to realize that the measurement needs of the teacher and the researcher differ. The teacher cannot rely entirely on quantifiable data. His task is to assess each student's progress toward specified instructional objectives. He does not necessarily need to compare students. Nor should he be constrained by assumptions about normal population distributions when he assigns grades.

The researcher, on the other hand, usually makes comparisons. His sampling procedures, his assignment of variables such as sex, age, or I.Q., and his statistical procedures all imply the *comparative* nature of his task; he is almost always comparing one student or one group of students with another. In addition, the researcher must be more precise than the teacher in analyzing data, whatever its form; and he must meet more conditions and observe more constraints in gathering data.

This monograph is written primarily for the researcher. It reviews a number of attempts to measure appreciation of literature. The measurements are grouped in two categories: 1) discriminations among poems or prose extracts and 2) content analysis. Following the review is an evaluation of the limitations and possibilities of these measures. The monograph concludes with specific recommendations for further research into the problem of measuring growth in appreciation of literature.

In an attempt to strengthen both research and teaching in the area of literary study, a more focused definition of the general term *appreciation* will be used here. The term *appreciation* is used in this review to mean the process of deciding literary merit. *Appreciating* is the act of recognizing literary merit. We can observe the outcome of appreciating when we see a reader choose an original poem of merit over a rewritten, inferior version of the same poem. Consequently, we can write verifiable performance objectives for *appreciating*, objectives like the following: given a poem (or story or essay) of merit and a rewritten, inferior version of the poem, the student will choose the poem of merit. However, we can only guess at what the process of appreciation itself is like, how it develops, and how it might be enhanced.

It will be helpful to consider the relation of *understanding* and *valuing* to appreciation, as it is defined here. Making a discriminating appreciation of a poem involves understanding, yet it is possible for a reader to recognize merit in a poem without fully understanding it. A reader might comprehend equally well the statements in an original poem of merit and an inferior version of the same poem and still be unable to choose the poem of merit. At the same time, another reader is able to choose the poem of merit even though he finds the poem unattractive and unappealing. Discriminating appreciation is still possible even though the reader feels little personal attraction for the poem because of its tone or style or theme.

Appreciation, then, is based on understanding and can be independent of valuing. It is an *aesthetic* process, involving the evaluation of separate facets of the work and concluding with an overall assessment of its literary merit. Not every discriminating choice is based on a scholarly assessment of *all* the facets of a work, however. For example, a reader may recognize the poem of merit by perceiving only the superiority of the diction.

The phrase "literature of merit" is used here to mean any work of literature which is honest, original, and powerful. It might have been written yesterday by a sixth grader in Harlem or Iowa City or centuries ago by a British poet or a Greek playwright.

Excluded from this review are the various measures of attitudes toward literature or toward specific works of literature, measures like the semantic differential, projective tests, and Thurstone and Likert scales. Any one of these might be a useful adjunct to a study of growth in appreciation of literature.

Also excluded from this review are the most common measures in literature, those that assess understanding, perception, or interpretation of a work. For example, the new test, "A Look at Literature," developed cooperatively by the National Council of Teachers of English and Educational Testing Service (Princeton, N.J.: ETS, 1969) claims to be a measure of appreciation as well as of "critical reading." Actually it is only a measure of perceiving and interpreting—and a good one. Another adequate measure in this same category is the "Ability to Interpret Literary Materials" subtest of the Iowa Tests of Educational Development. (Chicago: Science Research Associates, 1960.) In both of the above tests the student is asked to answer multiple-choice questions about a poem or short prose selection which is presumably unfamiliar to him. In virtually all other published literature tests, the student is merely asked to recall facts about the author, period, genre, or specific work. Equally inappropriate to this review are measures of pupil preferences and attitudes which have little relation to the pupil's ability to discriminate between good and bad in published literature.

Review of Attempts to Measure Appreciation

Discrimination among prose extracts and poems

The most common measure of appreciation over the years has been the test which requires the subject to discriminate among poems or among extracts from poems or prose. These measures claim their content validity either from the source of the selection or from expert opinion. If the selection comes from a recognized classic, then it is assumed to be a valid item for the test. If experts in literature agree that the selection is good literature, it is considered a valid item. Often the items on the test are submitted to literary experts, who are asked to rank-order them by quality. Their ranking then becomes the correct ranking, and the subject's score is determined by how closely his judgment matches the experts' judgment. With one exception, all of the studies described below utilize one or both of these sources of content validity.

Various measures of prose discrimination will be described first. In a study of the prose preferences of school children, ages nine to fourteen, Ballard (1914) used an extract from Sir Thomas Malory's *Morte d'Arthur* with three different versions of the same extract, all of which he wrote himself. These versions he called the florid, the plain, and the jocular.

Speer's (1929) lengthy study of appreciation of poetry, prose, and art used specimens of already-rated prose from ten composition scales in wide use in the schools at the time of his study. The final form of the test included 30 paired prose specimens. Speer did not claim much for the results of the test, saying that the test indicated "merely recognition of difference between good and bad, good and better, and poor and poorer specimens of prose . . . ." It did score the pupil on gross recognition of differences in prose of varied degrees of acceptability (Speer, 1929, p. 41). Speer found a split-half reliability of .78.

Claiming both source and expert opinion as sources of validity, Carroll (1932) devised a test of prose appreciation for high school students. The test consisted of 12 sets of four prose extracts--one from a recognized author, one from a book generally considered to be of poor quality, one from an escapist fiction selection found in romance or movie magazines, and one a mutilation. All the extracts in one set were on the same subject. Standardized on three thousand Minnesota high school students, the test was originally distributed by the Educational Test Bureau, Minneapolis, Minnesota. Carroll claimed a reliability coefficient of .71 for both the split-halves and retest methods. The test has been used occasionally over the years in correlation studies (Carroll, 1934; Schubert, 1953; Burton, 1952). Later versions of the test were standardized on junior high and college populations. It is now out of print.

For their study of literary appreciation Williams, Winter, and Woods (1938) constructed a variety of tests. On the Age Scale Test, subjects (girls 11-17) were asked to rank a set of 15 compositions on the subject of "school." In the prose part of the Ranking Method Test, subjects sorted, according to preference and then resorted in the manner of the Q-technique, short prose extracts of a wide variety of merit. In the prose part of the Paired Comparison Test, the subject chose the better of two sentences. In the Triple Comparison Test, subjects chose the best of three sentences; in each of three sections of this test, excellence depended on the sound of the sentence, the logical construction of the sentence, and the aptness of particular words. In the prose part of the Triple Comparison Test, subjects were asked to choose the best from among three short prose extracts—the best usually taken from the *Oxford Book of Prose*, the intermediate from "an author of an intermediate type," and the worst from popular magazines. The experimenters did not place much faith in the reliability coefficients because the separate sections of the tests were too short and "the alternative forms too imperfect." They found reliabilities ranging from .36 to .94 for the various tests.

Burton (1951) chose two published short stories, one of them considered good literature of artistic merit and the other considered superficial and artistically second-rate. The student read the two sto-

ries and then took a 20-item multiple-choice test designed to test his ability to "critically compare" the two stories. Burton found a reliability coefficient of .74 for two forms of the test. Another test constructed by Burton for the same study presented the student with summaries of 10 contemporary short stories of merit. The summaries stopped at a certain point, and the student then rated for quality three versions of the conclusion to the story. For this test Burton found a split-halves reliability of .83.

After considering carefully the tests devised by Carroll and by Williams, Winter and Woods, Harpin (1966) constructed a test consisting of matched pairs of extracts from novels. One of the pair had "literary merit" while the other did not. The extracts were not identified by author or source. The final form of the test had nine pairs and a section with four passages to be arranged in order of preference. He obtained a test-retest reliability of .75.

Measures of appreciation of poetry have been very much like those for appreciation of prose. A study by Abbott and Trabue (1921) reports a test constructed by rewriting and deliberately making worse a well-known poem of quality, or a stanza from one. They began with a good poem, like Frost's "House Fear," and revised it for three inferior versions—a sentimental version, a prosaic version, and a metrical version, the latter intended "to render the movement either entirely awkward or less fine and subtle than the original." The final two forms of the test each had 13 of these sets of four. The test was wholly unreliable for the elementary grades but had a reliability coefficient of .44 for high school students, .65 for college students, and .72 for graduate students in English.

Speer's test of recognition of merit in poetry claimed its validity from an elaborate process of judging by experts (1929). The test consisted of 36 items of two poems each, one rated high, the other rated low, by the judges. The subject made a choice between the two on each item. The coefficient of reliability (split-half) was .68.

For her study of the effect of creative work on aesthetic appreciation, Leopold (1933) used both the Abbott and Trabue test described above and two tests of her own construction. In one she

selected a stanza from a well-known poet and then rewrote it twice, "aiming at a less and a greater degree of inferiority to the original." In the other she selected a short passage of a few lines from a recognized author and then deliberately weakened only the images and epithets in an alternate version.

In their study already mentioned, Williams, Winter, and Woods (1938) were interested in measuring appreciation of poetry as well as prose. On the Ranking Method Test, they used the Q-technique for indicating preference by sorting and resorting poems. In the Paired Comparison Test, the child chose the best poetic lines from two alternatives. In the Triple Comparison Test, the child chose from among three possibilities—one from the *Oxford Book of Poetry*, one from "an author of intermediate type," and one from a popular magazine.

A study by Britton (1954) relied for its validity on the source of the poem. The poet was well-recognized and the poems "had something to communicate." Britton himself wrote counterfeit poems to go with these, poems which "had nothing to communicate." On the test, the subject was asked to arrange the eight true poems and the seven counterfeit poems in order of preference. Using the results of earlier factor analytic studies of poetic preference by Eysenck (1940), Britton chose two each of the eight true poems to represent the two bipolar factors in Eysenck's report, "simple-complex" and "abandoned-restrained." This complexity gives his study a degree of sophistication lacking in the other studies described above. He did not examine his test for reliability.

Still another test of the ability to judge merit in poetry is the Rigg *Poetry Judgment Test*, the only test of the kind under review available from a commercial test publisher (Bureau of Educational Research and Services, The University of Iowa, Iowa City, Iowa). The test was copyrighted in 1942. It consists of 40 short extracts of poetry (two to six lines) from "poets of established reputation," each extract paired with a parody of it "purposely made inferior in some respect." At the high school level the reliability coefficient for the two forms of the test is .84. The examiner's manual does not describe the subjects on whom the test was standardized. The author

of the examiner's manual notes that a high score on the test does not correlate highly with amount of instruction and then explains: "This conclusion is supported by the fact that one-fourth of the high school students who have taken this test do better than the average college student, and about six percent of these high school pupils score better than the lower fourth of the expert group, consisting to a large extent of college professors of English." The Rigg test has been used in correlation studies (Rigg, 1937) as well as in controlled experimental studies (Terrey, 1965).

Two final tests of appreciation should be mentioned. They are unique in that they remove from the original poem or prose extract a single word or a short phrase and then group the removed portion with two or three counterfeit portions, the student being asked to make his selection in the manner of a multiple-choice test. Fox (1938) removed two words from two spots in poetry extracts of about six lines and asked the subject to choose from among four phrases the right phrase for each spot. Eppel (1950) removed one line from a short poetry extract and then asked the student to select it from among two counterfeit versions of the same line.

Content analysis

Another way to assess appreciation of literature is by means of content analysis of the oral or written response. The response can be free, or it can be structured in reply to a set of specific questions. Although content analysis has been formalized only recently as a research tool, (Berelson, 1952; *Manual for Coders,* 1961) it has been in use informally for a long time. This section will note an early example of informal content analysis and will then review some significant recent research.

I.A. Richards' *Practical Criticism* (1929) is the classic analysis of the reading difficulties of critics of poetry, in this case college students. For many years Richards made a practice of asking his own students to write down their responses to poems which varied greatly in quality. Richards' book is a detailed report on these responses, and it continues to have great influence on studies of interpretation and

response and on the teaching of literature in schools and colleges. As he searched the responses looking for errors in interpretation and response, he found the following to be the main problems of his readers: 1) an inability to understand the poem as a statement or expression, 2) an inability to perceive the form of the poem and the movement and rhythm of the lines, 3) an inability to respond fully to imagery, 4) a tendency to be misled by erratic associations, 5) a reliance on stock responses, 6) a proneness to sentimentality and inhibition, 7) an unwillingness to judge the worth of poetry alone apart from the views and beliefs about the world it contained, and 8) an unwillingness to judge a poem for its own merits. These deficiencies are the categories of his analysis. Any one could prevent or distort an appreciation of literary merit.

Since the method of content analysis was formalized, several important studies of response to literature have appeared. One of the first and most important of these was a study reported by Taba (1955). One aspect of her year-long study was an examination of the extension of sensitivity by discussions. In order to code the 51 recorded class discussions of stories, she devised four categories: projections, generalizations, self-references, and irrelevancies. The categories are rather general; but within the first two there was a further breakdown into subcategories, six for "projections," and two for "generalizations."

A further development of this approach was a study by Squire (1964). He recorded the responses of 52 ninth and tenth grade students to four short stories. He studied these responses and then devised seven categories by which to code the elements of each student's responses: literary judgments, interpretational responses, narrational reactions, associational responses, self-involvement, prescriptive judgments, and miscellaneous. These same categories were used as a measuring instrument by Wilson in a controlled experimental study to assess the effects of classroom instruction and discussion on responses of college freshmen to three novels (Wilson, 1966; and also Sanders, 1970).

Purves (1966) attempted to devise a much more detailed and exhaustive set of categories and subcategories in his content-analysis study

of responses to literature. He studied the responses of literary critics, teachers, and students to works of literature and then devised a content analysis schema consisting of 120 elements grouped into four broad categories. The elements, worded to insure objectivity and not arranged into a taxonomy, are meant to describe any of the procedures or statements a writer uses in stating his responses to a work of literature. The elements, then, were derived from a close analysis of a large body of written material.

The categories, however, while intended to provide a useful and accurate way to cluster the elements, were devised primarily to indicate the postures or stances a responder can take toward a work. Looking in this way at the responders' relationship to the work, Purves identified four general relationships which became the categories of engagement-involvement, perception, interpretation, and evaluation. Of these, the one most directly relevant to assessing appreciation is *evaluation*. The coding within this category permits the identification of statements of either objective or subjective appraisal of a work. A survey of these statements in a student's written responses over the course of a year's work in literature would reveal any growth in his discriminative ability.

Synthesis

Discrimination among poems and prose extracts

It is clear from the review that discrimination measures have a long history. For both prose and poetry many different types of measures have been devised.

The most important questions to be asked of these discrimination measures are questions of validity. For a measure to be valid, it must give us information about the specified process or behavior.

The problem of content validity seems greatest in the *prose* discrimination measures. Since in some of the tests the prose extracts are very brief, the test might actually be measuring discrimination of stylistic features, rather than discriminative response to an entire short story or a whole novel. Most of the poetry discrimination tests, by contrast, offer choices between real poems and inferior versions of the same poems.

Measures of appreciation should also have face validity; that is, they should strike the student as reasonable and relevant tasks. The Rigg test has limitations here. The poetry extracts now seem rather old fashioned, and one wonders how adolescents these days would respond to them. Face validity may be lacking in an elementary school appreciation measure which contains poems more "adult" than those the students are familiar with. Furthermore, poetic styles and reader preferences change over the years. Another dimension of face validity is the personal preference individuals show for one poetry style or another, a preference illustrated by the simple-complex and abandoned-restrained bi-polar factors in Eysenck's study (1940).

A problem related to both content and face validity is the titling of appreciation measures. Until we know more about what we are

doing, it might be better to avoid general test titles such as "Poetic Appreciation Test." It would clearly be misleading to so label the Fox and Eppel tests, which remove words or lines from poems. The Fox test might best be called "Poetic Diction: A Test of Discrimination" and the Eppel test might best be labeled "Poetry Completion Test." The Carroll test would be more accurately titled "A Test of the Ability to Discriminate Among Prose Selections of Varying Quality."

Besides content and face validity, an appreciation measure should possess criterion-related, or predictive, validity. It should accurately predict the quality or level of appreciative and discriminative response to a literary work of the reader's own choice. It should also predict the quality of his choice of fiction: if he scores high on an appreciation measure, he should be able to choose fiction of high quality and artistic merit. Obviously, this kind of validity is difficult to ensure. One simple way to approach it would be to see how well a teacher's assessment of a student's appreciation of fiction correlates with his score on an appreciation measure. Another strategy which might be used would be to carefully examine the free reading choices of students scoring high and those scoring low on an appreciation measure.

Finally, an appreciation measure should ideally have construct validity. That is, it should really be a measure of the construct, "being an appreciative and discriminative reader of fiction." The concept of construct validity is a complex one in test construction, and a full discussion of it is inappropriate here. It would be useful to note, however, the two most common approaches to obtaining evidence of the construct validity of a measure. Both approaches have been used in the studies under discussion here.

One approach is to find out whether older students do better than younger students on the test. Carroll reported such data as evidence of the validity of his test. The other approach is to seek high correlations with similar tests. Burton reports such correlations of his two tests with Carroll's test. Since the correlations were rather high (.51 and .61), they provided some evidence of the construct validity of all the tests.

The attempt to establish construct validity actually raises a separate but very important empirical question: What is the nature of the construct "appreciation of literature"? Measures like the ones above can assist us in explaining and better defining it.

A measure that would satisfy all the conditions of validity can probably never be constructed. The task of the test-maker is to put together the most convincingly valid test he can manage. We can do much better than we have.

Content analysis

In just a short period of time, content analysis of response to literature has seen remarkable technical development. The Taba and Squire categories and the four clustering categories in the Purves study provide a variety of schemes for gross analysis of oral or written responses. The 120 elements in the Purves coding system make possible an exhaustively detailed analysis of responses.

This approach to measuring appreciation—and other aspects of response to fiction, as well—is very flexible. It can be based on an oral or written response. It can be obtained either in a test situation or in a natural situation, as in a tape recording of a small, student discussion group. It has the additional feature of being acceptable to researchers who doubt the validity of discrimination measures. The material for analysis is a student's own written or spoken essay of response to a literary work rather than the pattern of his choices on a discrimination test.

Content analysis is suited to assessing large-scale shifts in group patterns of response as a result of instruction. This makes it a useful research tool in studies of the effectiveness of instructional strategies and curriculum materials.

A disadvantage of content analysis is that it is time consuming and costly. Coding the separate statements in a set of essays takes many hours. Several hours are required to train an analyst to use a coding system like that in Purves' study.

Recommendations

We need reliable and valid measures of appreciation to serve four research purposes: 1) to enable us to trace the development of appreciation of literature from childhood into adulthood; 2) to permit us to test claims now being made for the effect on appreciation of certain materials and modes of instruction; 3) to permit us to test the efficacy of experimental programs aimed at enhancing growth in appreciation of literature; and 4) to help us deepen our understanding of the construct "appreciation of literature."

Specific recommendations

1. Researchers should be careful not to confuse tests of appreciation with reading comprehension tests or with literary tests of interpretation and understanding. Measured reading comprehension is not the same as the discriminative appreciation of literature under discussion in this monograph. If the two were the same, we would expect to find general reading tests highly correlated with appreciation measures. Instead, the correlations reported so far range from low to moderate: .47 with the Carroll test, .27 and .33 with the two Burton tests, and no significant difference on Carroll test scores between a group of retarded readers and a group of "unselected readers" (Schubert). Actually, the reported correlations of general intelligence with appreciation are higher: .54 with the Carroll test, .44 and .64 with the Burton tests.

Clearly, appreciation of literature is related to general reading comprehension and measured intelligence, but it is something more, as well. An appreciation hypothesis in a research study requires an appreciation measure like the ones under review here, not just a measure of understanding or of comprehension.

2. We need factor analytic studies using a variety of appreciation

measures. In factor analytic studies a large number of tests (usually related) are given to the same subjects. The test scores are then correlated and the large matrix that results can be examined for clusters of correlations that might yield definable "factors" of appreciation. Factor analysis can show which tests appear to be measuring the same things and thereby help us reduce duplication in assessing appreciation and reduce the number of variables the researcher needs to be concerned about.

Gunn (1951) found a general aesthetic factor and a "technical" bi-polar factor in his study of factors in the appreciation of poetry. Eysenck (1940) reported two bi-polar factors: emotional-restrained and simple-complex. Rees and Pederson (1965) identified six factors or "points of view" in the reading of poetry among college students. We need more studies of this type. With a better understanding of the "factors" involved in appreciating literature, we can design better measures of appreciation. Factor analysis contains no magic, as researchers in intelligence and reading have discovered. Considerable logical analysis and the selection of carefully refined test items are requisite to any meaningful factor analysis. Nevertheless, we have not yet tested its limits in appreciation studies.

3. Discrimination tests should be designed with a larger number of items, to enhance reliability. The Abbott and Trabue test, with only 13 items, had a reliability of only .44 in the high school and .65 in the college. By contrast, the Rigg test, with 40 items, reached a reliability of .84 in the high school. It is only generally true that more items mean higher reliability—Harpin achieved a reliability of .75 in the high school with only 10 items—but with the measures reviewed here the trend is for more items to yield higher reliability.

4. Discrimination tests should be designed to yield higher reliabilities in the elementary school. Several of these—Abbott and Trabue, for example—report virtually complete unreliability in the lower grades with increasing reliability through the secondary school and college. It is true that when scores increase with age on a test, we have evidence of the construct validity of the test. However, unless we want to assume that younger children are incapable of discrimina-

tion, we should work to devise measures that can be used with some reliability in studies of appreciation in the early grades. One approach would be to reduce reading dependency by reading the items aloud to the students or by playing an audio recording of the items. Another approach would be to make the test items more accessible to the students; for example, one could use good poetry written by children or for children, rather than "adult" poetry.

5. Measures of prose appreciation need to present choices between longer prose selections, as in the Harpin study, or between complete short stories, as in the two measures devised by Burton. Burton's test of choice between the two complete stories looks very useful and would probably have higher content validity than prose discrimination tests using only single sentences or too-short paragraphs. It could be extended to include several pairs of stories, perhaps eight or ten, with the student being asked to choose the better. Of course, administering such a test would take more time, but that could be kept within reasonable bounds if brief short stories were used and if they were presented on audio tape recordings, with students following on written scripts. The recorded voice could set the pace of the reading and make the exact time requirement of the test known in advance. In addition, the recording could control for differences in reading ability, a skill shown in Burton's study to have a rather low correlation with scores on measures of appreciation (.31 and .40).

6. In content analysis studies of written or oral response to specific works, greater use should be made of the Purves coding system. It is based on thorough research and analysis. Such a valuable research tool should be widely used. Furthermore, if several different researchers use the same coding system, results can be easily compared and collated and we could begin to accumulate knowledge about response to literature. Researchers should know that an appendix to the Purves study explains in detail how to train analysts and how to code and score a response. There is even a suggested format for a scoresheet.

7. Research studies with appreciation hypotheses should use *several* measures of appreciation, not just a single measure. In such studies the combined weight of several different, perhaps quite varied, appre-

ciation measures such as the ones under review here will nearly always be more convincing than a single measure (Webb, *et al.*, 1966).

8. This final recommendation places the problem of measuring appreciation, as it is defined in this review, in the context of the larger problem of designing a study of the effects of instruction in literature. In studying the effects of literary instruction in the classroom, we need to state larger sets of research hypotheses and use a variety of measures to test them. If a researcher is examining the effects of an experimental program of literary study, he could hypothesize various kinds of changes—understanding, interpretation, appreciation, attitude, even specific literary-critical skills—and use separate measures for each hypothesis. For *understanding* and *interpretation* he could use the ETS/NCTE test "A Look at Literature" or Andresen's "Literary Profundity Test" (Andresen, 1969) or the appropriate levels of a content analysis of spoken or written responses. For *appreciation* he could use tests such as those under review here. For *attitude* he could use any one of several types of attitude measurement—Thurstone, Likert, or semantic differential scales, to name just three. For specific *skills* in literary criticism he could use the Fox test for sensitivity to diction in poetry and prose; and he could construct additional tests of specific skills, like the ability to recognize both the vehicle and tenor of metaphor, for example.

New computer-based methods of multivariate analysis of variance (Bock and Haggard, 1968; Hoetker, 1971) make it possible for us to include in a single analysis several dependent variable scores from tests like those suggested above. In other words, we can now examine the differences between several groups—for instance, between three experimental groups and one control group—on several measures at the same time. Our studies should be at least as sophisticated as the best means of analysis.

This review and these recommendations do not imply that more informal methods of assessing appreciation, such as interviews, case studies, and shrewd observation, are inappropriate or unproductive. There is nothing sacrosanct about quantifiable data. In the rudimentary state of our knowledge about appreciative discrimination of literature, any convincing new information, whatever its form, is needed.

21

Bibliography

Abbott, Allan and Trabue, M. R. "A Measure of Ability to Judge Poetry," *Teachers College Record,* 22 (March 1921) 101-26.

Anastasi, Ann. *Psychological Testing* (New York: Macmillan, 1968, 3rd edition) 665 p.

Andresen, Oliver. "The Significance of Profundity in Literary Appreciation," *Reading Research Quarterly,* 5 (Fall 1969) 100-18.

Ballard, P. B. "Prose Preferences of School Children," *Journal of Educational Psychology,* 5 (1914) 10-21.

Berelson, Bernard. *Content Analysis in Communication Research.* (Glencoe, Illinois: The Free Press, 1952) 222 p.

Bock, Daryl R. and Haggard, Ernest A. "The Use of Multivariate Analysis of Variance in Behavioral Research," *Handbook of Measurement and Assessment in Behavioral Sciences.* D. K. Whitla, Ed. (Reading, Massachusetts: Addison-Wesley, 1968) pp. 100-42.

Burton, Dwight L. *A Comparison of Three Methods for Teaching Appreciation of the Short Story to Twelfth-Grade Students.* (Unpublished Doctoral Dissertation, University of Minnesota, 1951.)

Burton, Dwight L. "The Relationship of Literary Appreciation to Certain Measurable Factors," *Journal of Educational Psychology,* 42 (November 1952) 436-39.

Carroll, H. A. "A Standardized Test of Prose Appreciation for Senior High School Pupils," *Journal of Educational Psychology,* 23 (September 1932) 401-10.

Carroll, H. A. "Appreciation of Literature and Abstract Intelligence," *Journal of Educational Psychology,* 25 (January 1934) 54-57.

Educational Testing Service. *A Look at Literature.* Princeton, New Jersey, 1968.

Eppel, E. M. "A Test of Poetry Discrimination," *British Journal of Educational Psychology,* 20 (June 1950) 111-16.

Eysenck, H. J. "Some Factors in the Appreciation of Poetry and Their Relation to Temperamental Qualities," *Character and Personality* (now *Journal of Personality*), 9 (1940-41) 160-67.

Forehand, Garlie. "Problems of Measuring Response to Literature," *Clearing House,* 40 (February 1966) 369-75.

Fox, Charles. "The Method of Testing Literary Appreciation," *British Journal of Psychology,* 29 (July 1938) 1-11.

Gunn, Douglass G. "Factors in the Appreciation of Poetry," *British Journal of Educational Psychology,* 21 (June 1951) 96-104.

Harpin, W. S. "The Appreciation of Prose," *Educational Review,* 19 (November 1966) 13-22.

Hoetker, James. *Students as Audiences: An Experimental Study of the Relationship Between Classroom Study of Drama and Attendance at the Theatre.* (Urbana, Illinois: National Council of Teachers of English, 1971) 139 p.

Institute for Social Research. *Manual for Coders: Content Analysis at the Survey Research Center.* Ann Arbor, Michigan, 1961.

Iowa Tests of Educational Development. Chicago, Science Research Associates, 1960.

Kerlinger, Fred N. *Foundations of Behavioral Research.* (New York: Holt, Rinehart, and Winston, 1964) 739 p.

Leopold, Kathleen B. "The Effect of Creative Work on Aesthetic Appreciation: An Experiment in the Teaching of Poetry," *British Journal of Educational Psychology,* 3 (February 1933) 42-63.

Purves, Alan C. *Elements of Writing About a Literary Work: A Study of Response to Literature.* (Urbana, Illinois: National Council of Teachers of English, 1966) 90 p.

Rees, Richard D. and Peterson, Darhl M. "A Factorial Determination of Points of View in Poetic Evaluation and Their Relation to Various Determinants," *Psychological Reports,* 16 (1965) 31-39.

Richards, I. A. *Practical Criticism.* (New York: Harcourt, Brace, and World, 1929) 362 p.

Rigg, Melvin D. "The Relationship Between Discrimination in Music and Discrimination in Poetry," *Journal of Educational Psychology,* 28 (1937) 149-52.

Rigg, Melvin D. *The Rigg Poetry Judgment Test.* (Iowa City, Iowa: Bureau of Educational Research and Services, State University of Iowa, 1942).

Sanders, Peter L. *An Investigation of the Effects of Instruction in the Interpretation of Literature on the Responses of Adolescents to Selected Short Stories.* (Unpublished Doctoral Dissertation, Syracuse University, 1970.)

Schubert, Delwyn A. "The Relationship Between Reading Ability and Literary Appreciation," *California Journal of Educational Research,* 4 (May 1953) 201, 202.

Speer, Robert K. *Measurement of Appreciation in Poetry, Prose, and Art, and Studies in Appreciation.* (New York: Teachers College, Columbia University, Contributions to Education No. 362, 1929) 77 p.

Squire, James R. *The Responses of Adolescents While Reading Four Short Stories.* (Urbana, Illinois: National Council of Teachers of English, 1964) 65 p.

Taba, Hilda. *With Perspective on Human Relations.* (Washington, D. C.: American Council on Education, 1955.)

Terrey, John Naylor. *An Experiment Comparing Close Study and Wide Reading as Factors Contributing to an Ability to Judge Merit in Poetry.* (Unpublished Doctoral Dissertation, Washington State University, Pullman, 1965) 162 p.

Webb, Eugene J., Campbell, Donald T., Schwartz, Richard D. and Sechrest, Lee. *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* (Chicago: Rand McNally and Company, 1966) 225 p.

Williams, E. D., Winter, L., and Woods, I. M. "Tests of Literary Appreciation," *British Journal of Educational Psychology,* 8 (November 1938) 265-83.

Wilson, James R. *Responses of College Freshmen to Three Novels.* (Urbana, Illinois: National Council of Teachers of English, 1966) 42 p.

ERIC/CRIER+IRA Publications

The following publications are the cooperative products of ERIC/CRIER and IRA and are available from the International Reading Association, Six Tyre Avenue, Newark, Delaware 19711. (See page 29 for ordering microfiche.)

Reading Research Profiles Bibliography Series

Leo Fay. *Organization and Administration of School Reading Programs,* International Reading Association, Newark, Delaware, 1971, 64 p. Members' price $1.00, nonmembers' $1.50 from IRA.
ED 046 677, microfiche $0.65 from EDRS.

James L. Laffey. *Methods of Reading Instruction,* International Reading Association, Newark, Delaware, 1971, 87 p. Members' price $1.00, nonmembers' $1.50 from IRA.
ED 047 930, microfiche $0.65 from EDRS.

Roger Farr. *Measurement of Reading Achievement,* International Reading Association, Newark, Delaware, 1971, 96 p. Members' price $1.00, nonmembers' $1.50 from IRA.
ED 049 906, microfiche $0.65 from EDRS.

Leo Fay. *Reading Research: Methodology, Summaries, and Application,* International Reading Association, Newark, Delaware, 1971, 75 p. Members' price $1.00, nonmembers' $1.50 from IRA.
ED 049 023, microfiche $0.65 from EDRS.

Samuel Weintraub. *Auditory Perception and Deafness,* International Reading Association, Newark, Delaware, 1972, 51 p. Members' price $1.00, nonmenbers' $1.50 from IRA.

Samuel Weintraub. *Vision-Visual Discrimination,* International Reading Association, Newark, Delaware. In press.

20 Year Annotated Index to *The Reading Teacher*

Available from the International Reading Association for $3.00 to members of the Association and $3.50 to nonmembers. It is also available from EDRS (ED 031 608) in microfiche for $0.65.

Monographs

ERIC/CRIER+IRA monographs are published by the International Reading Association and are available to members of the Association at a special rate. The monographs are also available in microfiche from the ERIC Document Reproduction Service.

A. Sterl Artley. *Trends and Practices in Secondary School Reading,* International Reading Association, Newark, Delaware, 1968, 131 p. Members' price $3.00, nonmembers' $3.50 from IRA; ED 024 560, microfiche $0.65 from EDRS.

Ruth Strang. *Reading Diagnosis and Remediation,* International Reading Association, Newark, Delaware, 1968, 190 p. Members' price $3.00, nonmembers' $3.50 from IRA; ED 025 402, microfiche $0.65 from EDRS.

Roger Farr. *Reading: What Can Be Measured?* International Reading Association, Newark, Delaware, 1970, 299 p. Members' price $3.25, nonmembers' $4.95 from IRA; ED 032 258, microfiche $0.65 from EDRS.

James L. Laffey. *Reading in the Content Areas.* International Reading Association, Newark, Delaware, 1972, 236 p. Members' price $3.00, nonmembers' $4.50 from IRA.

Other Interpretive Papers

A series of interpretive papers on discrete topics in reading and aimed at specific audiences, such as teachers, includes:

Nicholas Anastasiow. *Oral Language: Expression of Thought.* International Reading Association, Newark, Delaware, 1971, 51 p. Members' price $1.00, nonmembers' $1.50.

Ivan Quandt. *Self-Concept and Reading.* International Reading Association, Newark, Delaware, 1972. In press.
Members' price $1.00, nonmembers' $1.50.

The following papers are available in the Reading Information Series: WHERE DO WE GO?

Eugene Jongsma. *The Cloze Procedure as a Teaching Technique.* International Reading Association, Newark, Delaware, 1971, 42 p.
Members' price $1.00, nonmembers' $1.50.

MaryAnne Hall. *The Language Experience Approach for the Culturally Disadvantaged.* International Reading Association, Newark, Delaware, 1972, 42 p.
Members' price $1.00, nonmembers' $1.50.

Diane Lapp. *The Use of Behavioral Objectives in Education.* International Reading Association, Newark, Delaware, 1972, 56 p.
Members' price $1.00, nonmembers' $1.50.

Lawrence G. Moburg. *Inservice Teacher Training in Reading.* International Reading Association, Newark, Delaware, 1972, 42 p.
Members' price $1.00, nonmembers' $1.50.

John Paul Jones. *Intersensory Transfer, Perceptual Shifting, Modal Preference, and Reading.* International Reading Association, Newark, Delaware, 1972.
Members' price $1.00, nonmembers' $1.50.

Charles R. Cooper. *Measuring Growth in Appreciation of Literature.* International Reading Association, Newark, Delaware, 1972.
Members' price $1.00, nonmembers' $1.50.

Micromonographs

Pamphlets in a series designed to answer parental questions on reading are available from IRA in single copies at $0.35 to members and $0.50 to nonmembers. Bulk order price is $0.20 each on orders of 100 or more copies.

Rosemary Winebrenner. *How can I get my teenager to read?* 12 p.

Norma Rogers. *What is reading readiness?* 16 p.

Norma Rogers. *Can I help my child get ready to read?* In press.

Norma Rogers. *What books and records can I buy for my preschooler?* In press.

Molly Ransbury. *How can I get my primary child to read?* In press.

Marcia Baghban. *How can I help my child learn to read English as a second language?* In press.

Don Eberly. *How does my child's vision affect his reading?* In press.

Eric documents are available from:

ERIC Document Reproduction Service
Leasco Information Products Company (LIPCO)
Post Office Drawer O
Bethesda, Maryland 20014

Please furnish the following information:

1. The accession number (ED number) of the desired document
2. The type of reproduction desired—microfiche
3. The number of copies being ordered
4. The method of payment—cash with order, deposit account, charge
   a. The book rate or library rate postage is in costs quoted
   b. The difference between book rate or library rate and first class or foreign postage (outside continental United States) rate will be billed at cost
   c. Payment must accompany orders totaling less than $10.00
5. Standing orders for microfiche cost .089 cents per fiche, special collection cost is .14 per fiche, back collection cost is .089 per fiche.

EDRS will provide information on charges and deposit accounts upon request. All questions should be directed to J. Peter Maucher, Manager of Institutional Sales, or E. Brien Lewis, Manager of Client Sales at LIPCO.

INTERNATIONAL
BOOK YEAR
1972

UNESCO